



Astera Labs Unveils Orion: A Compute-Agnostic Rack-Scale Architecture for Token-Efficient AI Factories

October 12, 2026

Modular, production-grade reference designs are designed to help neoclouds and enterprises quickly scale AI inference capacity up and across with up to 160 accelerators per pod

News Highlights:

- **Bring Your Own GPU/CPU:** Modular Orion reference architecture is designed to give customers the flexibility to choose accelerators and CPUs based on availability, lead time, cost, target applications, or workloads.
- **A faster-path to AI inference capacity:** Six production-grade reference designs—spanning rack, compute tray, fabric tray, memory tray, link board, and compute board—*help to* accelerate time-to-market by reducing integration complexity and testing cycles.
- **Fabric Efficient Token Economics:** Orion's high-radix [320-lane PCIe switch fabric](#) is designed to scale from 8 to 160 GPUs per pod and to help the same GPU extract more tokens-per-second across AI inference workloads.
- **An open, standards-based roadmap:** Astera Labs plans to extend Orion's roadmap to additional open interconnects, including UALink™, to help protect customers' infrastructure investment and strengthen their supply chain resilience.

SAN JOSE, Calif., Oct. 12, 2026 (GLOBE NEWSWIRE) -- Astera Labs, Inc. (Nasdaq: ALAB), a leader in semiconductor-based connectivity solutions for rack-scale AI infrastructure, today announced Orion: a modular rack-scale reference design with production-grade connectivity that allows customers to choose their own GPU and CPU. To be offered through Astera Labs' ODM and OEM partners, the Orion production-grade reference designs are expected to help enable neocloud and enterprise customers to deploy AI infrastructure faster and run it more efficiently, turning racks built on open standards into token-efficient AI factories.

For neoclouds and enterprises to economically serve inference at scale, they must secure sufficient accelerator and CPU capacity while delivering predictable, low-latency performance across increasingly complex AI infrastructure. But because they must commit to that capacity long before they know which accelerator will be available or affordable for their workloads, tying early investment to a single platform carries unnecessary cost and risk. At the same time, ODMs and OEMs are forced to repeatedly solve complex integration, validation, and software challenges for every new accelerator platform—slowing deployments when inference capacity is in critical demand.

Orion addresses these challenges with a production-grade reference architecture that decouples rack-scale designs from accelerator and CPU selection. Built on open standards to support industry-leading inference accelerators and CPU ecosystems, Orion is designed to enable operators to deploy infrastructure once and scale inference workloads across the most cost-effective and readily available compute resources—via PCIe plug-in cards—without redesigning racks as technology and market dynamics evolve. At the core of Orion is a high-performance, standards-based, scale-up fabric engineered to connect up to 160 accelerators per pod and deliver predictable, low-latency communication and expanded memory capacity across racks and pods. Orion's optimized PCIe connectivity and composable infrastructure are additionally designed to help accelerators extract more tokens-per-second across AI inference workloads. Orion is delivered as a complete modular rack-scale reference design—including system configurations for the compute tray, fabric tray, memory tray, link board, and compute board—all integrated with COSMOS Rack Manager and COSMOS Fabric Manager orchestration software for management, fleet-wide telemetry, diagnostics, and AI infrastructure intelligence.

"We're entering a new era of AI inference, where operators are judged on the tokens they deliver per dollar, not the hardware inside their racks," said Pravin Patel, vice president and general manager of the AI Platform Solutions Group at Astera Labs. "Orion gives AI inference cloud providers a production-grade, compute-agnostic architecture so they can quickly bring inference capacity online, provide outcome-based pricing for their customers, and adapt as the industry and compute landscape evolves."

Orion rack reference and system designs are on track for availability through Astera Labs' OEM and ODM partners beginning in Q1 2027. Astera Labs will showcase a powered Orion rack—featuring interchangeable, pre-validated compute trays with different CPU and accelerator configurations at the 2026 Open Compute Project (OCP) Global Summit, taking place October 12-15 in San Jose, California.

Additional Resources:

- [Orion Product Page](#): architecture overview, specifications, and reference design details on the complete Orion product family
- ["Compute-Agnostic Rack Design for the AI Factory: Introducing the Orion Rack-Scale Reference Architecture"](#), blog by Pravin Patel, vice president and general manager, AI Platform Solutions Group, Astera Labs

Frequently Asked Questions:

Q: Why do neoclouds and enterprises need flexibility to use different accelerators?

A: Inference workloads can run efficiently across a range of accelerators, but the newest, high-end accelerator capacity is often the hardest for neoclouds and enterprises to secure or afford. AI inference cloud providers want to mix and match compute so they can consistently deliver competitive tokens per dollar, regardless of which accelerator is easiest to secure at a given time.

Q: Who are the target customer segments for these reference designs?

A: Orion is designed primarily for neoclouds and enterprises building AI inference capacity who need a fast, cost-effective way to scale inference infrastructure.

Q: What differentiates Orion from existing solutions?

A: Orion stands apart on four fronts: 1/ it is compute-agnostic, so customers can pick the best CPU or GPU, based on availability, lead time, cost, target applications, or workloads; 2/ Orion helps shorten time to market with production-grade ODM/OEM partners; 3/ the optimized connectivity and composable infrastructure are designed to deliver efficient AI inference performance and boost tokens-per-second; and 4/ its standards-based interconnect gives customers greater supply chain flexibility and scalable PCIe solutions.

Q: Will Astera Labs be manufacturing these racks?

A: No. Astera Labs designs and validates the Orion reference architecture and its component sub-systems. Complete racks and individual systems are on track to be manufactured, licensed, and sold through OEM and ODM partners.

Q: Describe the customer experience in building and deploying an Orion rack.

A: The experience is designed to be straightforward: a customer selects the accelerators and CPUs it wants to run, pairs them with Orion's COSMOS software stack that is both ready to be deployed and customizable to fit customer's needs. From there, customers work with an Astera Labs OEM or ODM offering Orion-based systems to purchase a complete Orion rack or the individual systems that fit their build. Interested parties may connect with Astera Labs for commercial inquiries.

Q: What components make up the Orion architecture?

A: Orion is delivered as a family of validated systems that can be purchased together as a complete rack or individually: Orion-CT, the compute tray offering modular CPU/GPU/XPU compute-platform options; Orion-FT, the fabric tray connecting compute resources across scale-up and scale-out topologies; Orion-CB, the compute board delivering standards-based electrical connectivity across CPUs, accelerators, switches, memory, NICs, and storage; Orion-LB, the link board extending connectivity across boards, trays, racks, and serviceable cable paths; and Orion-MT, the memory tray expanding cache and data resources for AI systems. All components are managed through Orion-RM, which not only manages the hardware devices and the connectivity, but also controls the fabrics with advanced intelligence features. More detail on each system is available on the Orion Product Page.

Q: Why would a neocloud or enterprise choose Orion over a fully-loaded, single-vendor rack?

A: The highest-performing single-vendor racks deliver excellent performance, but often at a steep capital cost, constrained supply, and with more memory or specialized hardware. These are excellent solutions for specific use cases like high performance computing or semiconductor design, but Orion is purpose-built for enterprise inference economics: it is designed to let inference providers pick accelerators based on cost, availability, and optimized workload, and deploy with significantly lower upfront capex than a fully-loaded single-vendor rack, and move toward outcomes-based, dollar-per-token pricing rather than committing to a single, fixed hardware configuration.

About Astera Labs

Astera Labs (NASDAQ: ALAB) provides rack-scale AI infrastructure through purpose-built connectivity solutions. By collaborating with hyperscalers and ecosystem partners, Astera Labs enables organizations to unlock the full potential of modern AI. Astera Labs' Intelligent Connectivity Platform integrates CXL®, Ethernet, NVLink Fusion, PCIe®, and UALink™ semiconductor-based technologies with the company's COSMOS software suite to unify diverse components into cohesive, flexible systems that deliver end-to-end scale-up, and scale-out connectivity. The company's custom connectivity solutions business complements its standards-based portfolio, enabling customers to deploy tailored architectures to meet their unique infrastructure requirements. Discover more at www.asteralabs.com.

Forward-Looking Statement Disclaimer

This communication, including the Frequently Asked Questions, contains forward-looking statements within the meaning of the federal securities laws. These include statements about the expected capabilities, performance, scalability, benefits, availability, customer experience, cost savings, customer adoption, commercial success and market opportunity of Astera Labs' Orion rack-scale reference architecture and related systems. They also include statements about the manufacture, licensing, sale and offering of Orion-based racks and systems through OEM and ODM partners, product roadmap plans, pricing models, and market trends and conditions (including a new era of AI inference). Forward-looking statements may be identified by words such as "allows," "boost," "can," "designed to," "engineered to," "expected," "help," "may," "move toward," "on track," "plans," "ready to be deployed," "to be offered," "will" and similar expressions.

These forward-looking statements are based on current expectations and assumptions. They involve risks and uncertainties that could cause actual results to differ materially from those expressed or implied by such statements. These risks and uncertainties include, among others, the risk that:

- demand for AI inference capacity, compute-agnostic rack-scale architectures and related AI infrastructure does not develop

as anticipated;

- customers do not adopt Orion, PCIe-based scale-up architectures, UALink or other open interconnects on the expected timelines or at all;
- Astera Labs' products do not achieve the expected benefits, including expected performance, interoperability, cost savings or tokens-per-dollar economics, in customer environments;
- testing, validation or simulation results are not indicative of real-world performance;
- the COSMOS software suite, including COSMOS Rack Manager and COSMOS Fabric Manager, does not deliver the expected management, telemetry, diagnostic or fabric intelligence capabilities across the Orion architecture;
- OEM and ODM partners do not manufacture, license, sell or support Orion-based racks and systems on expected timelines, including availability beginning in Q1 2027, or at all;
- customer evaluations, commercial inquiries, demonstrations or ecosystem collaborations do not result in production deployments or revenue;
- Astera Labs and its ecosystem partners are unable to develop, validate, manufacture, deploy or support products and technologies on expected timelines or at expected cost;
- third-party accelerators, CPUs, memory or other components are not available on expected timelines or at expected cost;
- technology, customer requirements, pricing models, competitive offerings or industry standards change;
- supply chain constraints or disruptions occur;
- litigation or disputes arise relating to Astera Labs' products or technologies;
- global economic, political and industry conditions, including those affecting the semiconductor and AI infrastructure markets, change;
- regulatory developments or international conflicts occur; and
- the other risks and uncertainties described in Astera Labs' Annual Report on Form 10-K, Quarterly Reports on Form 10-Q and other filings with the Securities and Exchange Commission materialize.

Forward-looking statements speak only as of the date they are made. Readers are cautioned not to place undue reliance on forward-looking statements. Astera Labs undertakes no obligation to update or revise any forward-looking statement, whether as a result of new information, future events or otherwise, except as required by law.

Media Contact:

Peter Lo

peter.lo@asteralabs.com

